Evolving Methods in Genetic Epidemiology
III. Gene-Environment Interaction in Epidemiologic Research

Quanhe Yang Ph.D. and Muin J. Khoury M.D., Ph.D.

I. Introduction

Genetic epidemiology is increasingly focused on the study of common diseases with both genetic and environmental determinants. The concept of gene-environment interaction is becoming a central theme in epidemiologic studies that assess causes of human disease in populations (1). Advances in genetic technology and the work of the Human Genome Project will make it easier for the study of gene-environment interaction to become an integral part of epidemiologic research. In this paper, we review epidemiologic concepts and definitions applied to gene-environment interaction and give an overview of both traditional and emerging approaches to the study of gene-environment interaction in epidemiologic research.

II. Concepts and Measurement of Interaction in Epidemiology

A. Evolving epidemiologic concepts and definitions of interaction

Over the last two decades, there has been much discussion about how to define and measure interaction in epidemiologic studies (2-12). Much of this discussion focused on deriving the expression for the relative risk for disease associated with exposure to multiple factors when the joint effects of these factors act through the same pathogenetic pathway. Two major definitions of interaction exist: statistical and biological (epidemiologic) interaction. From statistical perspective, the interaction of two or more risk factors is simply the coefficient of product term of these risk factors. Interaction is thus measured in terms of a departure from a multiplicative model (27). The use of statistical interaction has several advantages: it has convenient statistical properties; it has the ability to assess the extent of unknown confounding or bias; and it is easy to find a parsimonious models by keeping statistical interactions to a minimum (10). However, this method of measuring interaction has been criticized as ignoring any consideration of what constitutes interaction or synergy on the biological level, and as being inherently arbitrary and model-dependent (7-8, 12).

In biological interaction model, interaction between two factors is defined as their coparticipation in the same causal mechanism to the disease development (3, 8). Interaction is measured in terms of a departure from an additive model (8). In this model, at an individual level, a causal interaction effect can be understood by a hypothetical contrast of the outcome of a single subject under different exposure conditions to develop a disease. For example,

assuming two dichotomous risk factors (A and B) of a disease, a person would develop the disease at age 70 if exposed to A only, at age 60 if exposed to B only, and at age of 50 if exposed to both risk factors. The portion of the advance from 60 years of age to 50 years of age is called the interaction effect of A and B exposure. At the population level, if two factors can cause a disease, some cases of disease will involve exposure to both risk factors. In the absence of either factors, these cases would not occur (12). Several measurements and their confidence intervals were developed to measure the departure from an additive model: relative excess risk due to interaction, and attributable proportion due to interaction (8, 13-14). Studies also suggested that assessing interaction as departure from additivity is useful in assessing the public health implications in diseases prevention and in individual decision making in considering exposure to certain risk factors, such as smoking and alcohol (15).

B. Gene-environment interaction

There is accumulating evidence that allelic variations of many gene loci may play important roles in determining individual susceptibility to cancer (16-20) and other chronic diseases (21-23). In assessing the role of susceptibility alleles in disease risk, one should consider the effects of gene-environment interaction in disease etiology. Gene-environment interaction may be measured by the different effect of an exposure on disease risk among individuals with different genotypes or by the different effect of a genotype on disease risk among individuals with different exposures (24-25).

The concept of gene-environment interaction has long been recognized by geneticists (26), and occupied an essential place in ecogenetic studies which examine the genetically determined differences among individuals in their susceptibility to environmental risk factors (27-29). In recent years, an epidemiologic framework for evaluating gene-environment interaction has been proposed (24, 30-32). In a simple gene-environment interaction model, in which both the susceptibility genotype at a single locus and the environment exposure are considered dichotomous, one can construct an extended 2-by-2 table incorporating genetic and environment factors in studying disease etiology (24). Table 1 shows a simple gene-environment interaction model in the context of epidemiologic studies. For this simple model, it is assumed that unexposed individuals without the susceptibility genotype have a certain background risk for disease I. R_e refers to the relative risk for disease among people without the susceptibility genotype for disease who are exposed to the environmental risk factor relative to those with neither the susceptibility genotype nor exposure. R_g refers to the relative risk among people with the susceptibility genotype who are not exposed to the environmental risk factor relative to those with neither the susceptibility genotype nor exposure. R_{ge} is the ratio of disease risk among exposed people with susceptibility genotype to disease risk among unexposed people without the susceptibility genotype. This ratio reflects the strength of the gene-environment interaction.

Based on this simple gene-environment interaction model, the effects of six biologically plausible patterns of interaction on the relative risk of disease has been proposed (24) (Table 2). In type 1 interaction, the increased risk of

diseases was only observed when both genetic and environmental factors copartic ipate in the same pathogenetic mechanism, either the genotype alone nor the exposure alone causes excess risk (i.e., $R_g = R_e = 1$). In type 2 interaction, environmental exposure increases risk in individual without the corresponding genotype. In type 3 interaction, the genotype ($R_g > 1$) is associated with increased disease risk, whereas the exposure alone is not. In type 4 interaction, both the genotype and the environmental exposure are each associated with excess risk of disease ($R_g > 1$, $R_e > 1$). Type 5 and 6 interaction occur when there is a reversal of the genotype=s effect, depending on the presence or absence of the environment. In this case, the genotype is protective in the absence of the environment ($R_g < 1$), but is deleterious in the presence of the environment ($R_{ge} > 1$). Ottman (31) also proposed a similar model of studying gene-environment interaction in etiology of disease. This is a simplified gene-environment interaction model, the effects of gene-environment interaction on the measured phenotype are further complicated by the number of genetic loci involved and multiple environmental exposure factors, the moderation of the genetic effects, the dose of the environmental exposure, and the presence of etiologic heterogeneity (24, 31). Most studies have evaluated gene-environment interaction in terms of the departure from those predicted by the multiplicative model (33-34). Some investigators have suggested that many biologically plausible modes of gene-environment interaction involve extreme departures from multiplicative effects (35). For example, neither phenylalanine hydroxylase deficiency alone nor exposure to phenylalanine in the diet cause phenylketonuria (PKU); both must be present for PKU to develop (24). The gene-environment interaction may also be evaluated in terms of the departure from those predicted by the additive model.

III. Gene-Environment Interaction in Traditional Epidemiologic Studies

A. Strategies

The main emphasis of gene-environment interaction studies is not to localize the disease susceptibility genes or to find the inheritance patterns of the diseases, but rather to better understand the etiology and pathogenesis of the diseases through quantitative assessment of diseases risks in various populations (24, 31-32, 36-37).

Two types of genetic markers are used in gene-environment interaction studies: markers based on direct analysis of the DNA, and markers based on gene products such as specific blood groups, HLA antigens, serum proteins, and enzyme systems. When genetic markers are not available, family history data are sometimes used as a rough indicator of genetic susceptibility, though there is a potential for significant misclassification in using family history data in genetic epidemiologic studies (38-39).

With rapid advances and progress in molecular genetic technology and human genome project, the number of genetic markers and polymorphisms for all genes in human available for research will increase rapidly in the near future. The studies of gene-environment interactions are most meaningful when applied to functionally significant

variations in candidate genes which have a clear biological relation to or suspected of playing some role in the pathogenesis of disease (40-41).

B. Study design

If one views the gene-environment interaction as the genetic control of sensitivity to the environmental exposure, and genetic factors are regarded as one of the host characteristics, then gene-environment interaction can be analyzed through the use of the traditional epidemiologic study design: cohort, cross-sectional, and case-control studies.

When a relatively high number of polymorphic markers are located close to candidate gene loci, the case-control approach is a popular and effective means by which to study differences in genetic susceptibility and gene-environment interaction (24, 33). In a case-control design, the genetic markers and relevant environmental risk factors are each examined as independent predictors of disease and as interacting factors with the environmental exposures. The odds ratio of gene-environment interaction (R_{ge}) can be calculated as shown in Table 1. Examples of recent case-control studies include a study of interaction effects between maternal cigarette smoking and a transforming growth factor alpha (TGFA) polymorphism and the risk of oral clefts (42). The odds ratios for the exposure to smoking alone, or the TGFA genotype alone are close to unity, whereas the combined odds ratio for smoking and the genotype is 5.5 (95% C.I. 2.1-14.6), indicating evidence of gene-environment interaction for risk of oral clefts in offspring (42).

In a cohort study design, the environmental exposures and genetic risk factors are measured for all subjects at the start of follow-up (baseline) and possibly during follow-up. Despite of some major strengths of cohort study design (disease occurs or is detected after subjects are selected, and minimized selection bias), few cohort studies used genetic markers to test for effects of gene-environment interaction in disease etiology. It is partly due to the fact that the rapid development of molecular techniques are only seen recently and the mainstream of genetic analysis are to find the disease susceptibility genes. With the advances in molecular techniques and the findings of more candidate genes, one would expect to see increasing number of cohort studies to examine gene-environment interaction.

In cross-sectional design, the investigators randomly sample a set of individuals from a study population through a single ascertainment of disease prevalence. Individuals with different genetic and environment risk characteristics are compared with respect to the prevalence of the condition, and gene-environment interaction can also be tested (24). An example is the cross-sectional WHO-cardiac study of gene-environment in hypertension, stroke and atherosclerosis (43). Although cross-sectional designs are less time-consuming and able to examine many exposures and disease in the same study, the limitation of cross-sectional design for making causal inferences made its design less popular in the study of gene-environment interaction.

A number of case-control studies are including a familial component, for example, a family history of the

disease studied. The designs and some problems of the case-control studies incorporating family history are discussed in the epidemiologic literature (1, 34, 44). The study of familial aggregation in case-control studies can be extended by incorporating environmental covariates and their interaction with family history (45).

C. Choice of Controls: Population vs Families

In assessing gene-environment interaction, investigators can select control subjects either from the general population or from families, depending on the purpose of the study. If the investigators are assessing the prevalence of disease susceptibility genotypes in the general population and examining the interactions of those genotypes with environmental exposures for the risk of a disease concerned, investigators should use a population-based study design to choose control subjects.

Investigators assessing familial aggregation of a disease, evaluating whether such aggregation is caused by the presence of gene-environment interaction, should select control subjects from family-based study designs. Because the purpose of the study is not to make inferences to the general population, but to examine the familial aggregation of a disease. The family members are the only appropriate control subjects which will provide relevant information for the purpose of study (1).

D. Methodologic Issues in assessing gene-environment interaction

Mis-specification

In the presence of gene-environment interaction, quantifying the main effects of environmental factor alone or genetic factor alone can lead to mis-specification of the study model, and may miss important clues to the etiology of disease (46).

Errors of environmental exposure measurement

Precise measurement of an individual=s exposure to environmental risk factors are shown to be difficulty because of the individual=s ignorance of previous opportunity for exposure, the complex pattern of most long-term exposures, the lack of good biological indicators of exposure levels, and the lack of sufficient sources to collect individual exposure data on large populations (45). In the study of gene-environment interaction, the consequences of environmental exposure mismeasurement can lead to bias in the estimation of interaction effects and possible loss of precision and power with which interaction effects are estimated (24). Nondifferential misclasification is usually biased toward the null value, and differential misclassification may produce biased results in either direction. In addition to the errors of environmental exposure measurement, the timing of exposure during a developmentally important window is also important in examining gene-environment interaction. For example, the timing of the exposure to environmental exposure during the pregnancy and the development of a birth defect for a

genetically susceptible fetus.

Genotype misclassification

When measuring individuals' genotypes at the DNA level, misclassification can occur because of linkage disequilibrium (24, 47). Until a comprehensive catalog of common variants of all genes is developed, investigators must rely on genetic markers in the region of the candidate genes or in a nonexpressed portion of the genes in order to conduct many DNA marker-disease association studies. Under these circumstances, the observed differences in prevalence of a marker allele between case and comparison groups could be a result of linkage disequilibrium unless the actual sites of a deleterious variation involved in the disease are targeted (24, 48-49). Under linkage disequilibrium, Nondifferential misclassification can occur, and this misclassification may bias estimates of relative risk toward the null (i.e. OR = 1). Individual genotypes can also be measured by indirect methods. For example, some investigators used dapsone loading followed by urinary measurements of different metabolites to classify subjects as slow or fast acetylators in a case-control study of bladder cancer (50-51). Such indirect measures can lead to misclassification of the underlying genotypes of individuals. This type of misclassification is often independent and nondifferential. However, the argument that independent and nondifferential measurement errors produced bias only toward the null may not apply to assessments of gene-environment interaction. As with all types of interactions, independent and nondifferential misclassification may bias interaction estimates in any direction (12). Occasionally, genotype misclassification may be differential if the measurement method is affected by disease status itself or if a near-by gene is associated with the disease; such differential misclassification will further complicate the assessment of gene-environment interaction (1).

Confounding

Confounding is a major problem in evaluating gene-environment interaction. It can involve population subgroups with different genetic markers and disease frequencies. Unmeasured genetic determinants and environmental exposures can each act as confounders that could produce spurious associations. Race or ethnicity is an important source of confounding in studies of gene-environment interaction (52). One example is the reported association between the genetic marker Gm3;5;13;14 and non-insulin-dependent diabetes mellitus among the Pima Indians (53). In a cross-sectional study of this association, individuals with the genetic marker Gm3;5;13;14 were found to have a higher prevalence ratio of the disease than those without the marker (29% vs. 8%). This marker, however, turned out to be an index of white admixture. When the subjects of the analysis were stratified by degree of admixture, the higher prevalence of diabetes associated with the marker disappeared.

Confounding of interaction and dose-response

In traditional epidemiologic studies, dose-response relations refer to the changes in risk produced by changes in a single exposure, and interaction refers to changes in risk produced by two or more exposures. Dose-response relations and interaction may tend to confound one another (54). In assessments of the effect of gene-environment interaction on disease risk, the risk in disease associated with a certain genotype may vary depending on the environmental exposure, or the risk may be restricted to exposed persons only. Similarly, the effects of environmental exposures may vary depending on the genotype of the exposed person (25). For example, people who are slow acetylators of N-acetyltransferase 2 (NAT2) have an increased risk for bladder cancer, and the risk for bladder cancer associated with smoking may vary by NAT2 status (55). For slow acetylators of NAT2, current smoking and smoking in the distant past increased breast cancer risk in a dose-dependent manner. Those in the highest quartile (heavy smokers in the study) of cigarettes smoked 2 years previously were 4.4 (95% CI, 1.3-14.8) times more likely to develop breast cancer than those who never smoked (56). Failure to adequately model dose-response relations can lead to bias in gene-environment interaction estimates.

Sample size requirements for measuring gene-environment interaction

In an epidemiologic study of a given sample size, the power to detect statistical interactions is less than the power to detect main effects, and the variance of the interaction estimate will also be greater than the variance of the main effects estimate under a no-interaction model (7, 57-58). Several investigators examined the sample size and power calculation needed to detect gene-environment interaction in case-control studies (59-61). The data needed to calculate the sample size required to detect gene-environment interaction can be shown by a 2-by-4 table as is done in Table 3. This table lists six parameters: 1) The odds ratio of interaction (R_{ge}); 2) The odds ratio of having the disease among exposed individuals without the susceptible genotype relative to those with neither the susceptibility genotype nor exposure (R_e); 3) the odds ratio of having the disease among people with susceptible genotype but without environmental exposure relative to those with neither the susceptibility genotype nor exposure (R_g); 4) the prevalence of exposure in the population (e); 5) the prevalence of the genotype in the population (g); 6) the case/control ratio (59-60). The results of several studies have suggested that when the frequency of exposure is not extremely low or high, and the susceptible genotype is common, a modest sample size will be adequate to detect gene-environment interaction. For example, when the frequency of exposure and the prevalence of the genotype both range between 30% to 70%, about 200 case subjects and 400 control subjects (for case/control ratio 1:2) should be adequate to detect an odds ratio of gene-environment interaction (R_{ge}) greater than 4 with 80% statistical power (60). However, the susceptible genotypes for many common diseases are relatively rare, with prevalence ranging from 1 to 5%, and both the genotype alone (R_g) and exposure alone (R_e) have moderate effects on risk for disease. For example, the frequency of the BRCA1 185delAG among Ashkenazi Jews (62) is about 1%, and the odds ratios for BRCA1 (R_g) is about 2

(38, 63); therefore, a relatively large number of case and control subjects are needed to detect gene-environment interaction (usually more than 1,000 cases) (60). With such diseases, alternative approaches to detecting gene-environment interaction may be needed. These approaches include 2-tier sampling strategies (64-65), family or sibling-based designs (61), and case-only designs (66).

IV Gene-environment Interaction in Nontraditional Epidemiologic Studies

Concerns about selecting appropriate control subjects for case-control studies have led to the development of several nontraditional approaches in the study of genetic factors in disease (1, 34). These approaches involve the use of an internal control group rather than an external one. We will review three of these nontraditional approaches in detecting gene-environment interaction: 1) the case-only study, 2) the case-parental control study, and 3) the affected relative-pair study. Except for the case-only design, these nontraditional approaches were not developed with the intention of evaluating gene-environment interaction. Table 4 summarizes the features of these studies, including their assumptions, strengths, and limitations. We also briefly review use of the twin study to evaluate gene-environment interaction.

A. Case-only studies

The case-only design has been promoted as an efficient and valid approach to screening for gene-environment interaction under the assumption of independence between exposure and genotype in the population (67-68). If one's primary interest is in assessing possible interaction between genetic and environmental factors in the etiology of a disease, one may do so without employing control subjects. The basic set up for a case-only design is a 2-by-2 table (Table 5). The odds ratio calculated from a case-only design is related to the odds ratios for the exposure alone, the genotype alone, and their joint effects in the case-control design by the following formula:

$$OR_{ca} = R_{ge}/(R_e * R_g) * OR_{co},$$

where OR_{ca} is the case-only odds ratio, and OR_{co} is the odds ratio among control subjects relating the exposure and the susceptibility genotype. Assuming independence between the genotype and the exposure in the population, the expected value of OR_{co} becomes unity, and the odds ratio obtained from a case-only study measures the departure from the multiplicative joint effect of the genotype and the exposure. Under the null hypothesis, $OR_{ca} = 1$; $OR_{ca} > 1$ if the joint effect is more than multiplicative; and $OR_{ca} < 1$ if the joint effect is less than multiplicative (e.g., additive) (34). Confidence intervals of case-only odds ratio can be obtained by using standard crude analyses or logistic models that control for the effects of other covariates.

Table 6 shows data from a case-control study of the association between cleft palate, maternal smoking and TGFA polymorphism derived from Hwang et al. (42). The case-only OR_{ca} of 5.1 (95% CI, 1.5-18.5) calculated from Hwang et al. (42) can be compared with the odds ratio of the interaction 5.5 (95% CI 2.1-14.6) derived from their case-control study. Both odds ratios suggest a significant interaction between TGFA polymorphism and maternal smoking in the risk for cleft palate among the offspring. Study has shown that the case-only design requires fewer case subjects than case-control design to detect gene-environment interaction (66).

In applying the case-only design to test gene-environment interaction, investigators assume independence of the distribution of exposure and genotype in the population. This assumption may seem reasonable for a wide variety of genes and exposures, but there are some genes whose presence may be associated with a higher or lower likelihood of the exposure on the basis of some biologic mechanisms (34). The gene-environment interaction (OR_{ca}) derived from a case-only design assumes a departure from multiplicative effects. Studies have shown that many biologically plausible modes of gene-environment interaction involve a departure from multiplicative effects (35). If the true underlying model of joint effect is additive, the odds ratio of interaction (OR_{ca}) derived from a case-only design is questionable.

B. Case-parental control studies

The case-parental design may be an effective method of dealing with the effects of confounding by population stratification (69-71). In addition, when disease alleles are common and have modest effects, an association study may provide a more sensitive test for linkage between genetic markers and disease susceptibility genes than the classical linkage analysis (41). Several methods (72-77) combine the advantages of linkage and population association analyses and also take into account the effect of confounding. All these methods consider the alleles found in the parents of an affected offspring and compare transmitted and untransmitted alleles of parents to the affected offspring (transmission/disequilibrium test). Investigators using these methods can compare the genotype of the affected offspring with the genotype of a fictitious control subject carrying the nontransmitted alleles from each parent. The 2-by-2 table used in such a comparison is shown in Table 7. Odds ratios can be calculated in an analysis following that of a matched-pair design (34). To test gene-environment interaction, investigators can stratify case subjects according to their environmental exposure status (presence or absence) and can use the difference of odds ratios derived with and without the environmental exposure as an indication of departure from multiplicative interaction (34).

One limitation of this method could be that the Acontrol@ group may not be representative of the underlying population at risk, especially when certain parental genotypes associated with disease status may interfere with reproduction. In other study (78), investigators proposed using a noniterative method, which compares risk among those with a specific genotype with the risk among those with a comparison genotype. To study gene-environment

interaction, investigators can stratify on the environmental factor to obtain stratum-specific estimates of the disease-gene association, and the difference in the stratum-specific estimates reflect gene-environment interaction (78).

The need for the parents of the case subjects to be genotyped is another limitation of case-parental approach. The parental marker data may not be available for some case subjects, especially in studies of the genetic etiology of diseases with older age at onset. In other studies, investigators developed a method using marker information on all members of a nuclear family to infer the probability distribution of missing parental marker data (79).

C. Affected relative-pair studies

The third type of nontraditional epidemiologic method that can be used to test gene-environment interaction is the affected sib-pair or affected relative-pair method (80-84). In sib-pair analysis, investigators determine whether each sib-pair shares 0, 1, or 2 alleles identical by descent (IBD) at a locus of interest. Under random segregation, the expected distribution of sharing 0, 1, or 2 alleles is 25%-50%-25% between two siblings IBD. Departure from this distribution suggests linkage between the disease and the marker locus (84).

In contrast to the case-only and case-parental approaches, the sib-pair method is primarily used to test for genetic linkage when the genetic model underlying the disease is not known, especially for the diseases involving complex traits (1). The sib-pair methods can be incorporated into family-based epidemiologic studies (cohort and case-control designs): such incorporation allows investigators to control for suspected nongenetic risk factors and to test for gene-environment interaction in searching for genetic linkage (85-86). To look for gene-environment interaction using this method, investigators can stratify the affected individuals by their exposure status or incorporate the gene-environment interaction term in a multivariate analysis. For example, they can use logistic regression when testing for genetic linkage (86-87). The basic set-up for analyzing gene-environment interaction through sib-pair analysis is shown in Table 8. The difference of odds ratios for diseases between exposed and unexposed individuals are taken as an indication of gene-environment interaction.

The sib-pair method requires families with at least one affected member in addition to the proband. This requirement restricts the number of families for which this analytic method can be used. Because the affected relative-pair approach assumes Mendelian transmissions for expected distributions, any departure from independent segregation and random assortment could affect the results. Finally, selection factors, including survival, chronicity, and method of case ascertainment, may substantially affect the types of case subjects that could be available for this analysis (78, 86).

D. Twin studies in gene-environment interaction

The premise behind twin studies is that, because monozygotic twins (MZ) have 100% of their genes in common whereas dizygotic twins (DZ) have only 50% of their genes in common, an excess disease concordance

among MZ twins may reflect a greater role of genetic factors. Several investigators have extended the classical twin study to test for gene-environment interaction (25, 88-89). For example, Ottman (25) developed a method to test for gene-environment interaction on disease risk conditional on twin exposure status and genotype. This method involved two measures of relative risk: 1) relative risk for disease among exposed vs. unexposed cotwins, stratified by zygosity and proband exposure status (RR_e), and 2) relative risk for disease among MZ vs DZ cotwins, stratified by exposure status of the proband and cotwin (RR_z). Ottman then examined the behavior of the two measures under different assumptions about the relative effect of exposure and genotype on disease. RR_e reflects the effect of exposure on disease risk. When gene-environment interaction is present, RR_e is expected to differ between MZ and DZ twins because of their different probabilities of having the high risk genotype. RR_z reflects the effect of genotype on disease risk. When gene-environment interaction is present, RR_z is expected to differ between exposed and unexposed twins. In another study, investigators used a case-control design to calculate the odds ratios for disease among affected vs unaffected cotwins and compared these odds ratios among the various strata defined by exposure in the index twin. Gene-environment interaction is indicated by the difference in odds ratios by stratified environmental exposures (89). Recently, other investigators extended the twin study method by including the half-siblings in a study of genetic and environmental factors in the etiology of disease (90). Given the possible confounding by shared environmental factors (intrauterine and postnatal) and selection factors, the effects of gene-environment interaction obtained from twin studies should be interpreted with caution (1).

V. Some recent Developments

A. Linkage vs association studies

Recently, investigators argued that traditional linkage analysis has limited power to detect genes with modest effects, and suggested that association studies (Case-parental and affected sib-pair studies described in this review are forms of association studies) have more statistical power to detect genes of modest effect (41, 91). The key limitation for association study is that the actual gene or genes involved in the disease must be tentatively identified before the analysis can be carried out (41). However, with the rapid development of Human Genome Project and identification of major variants of human genes, association studies may become important methods of the future genetic analysis of complex traits and gene-environment interaction.

B. Population-based family study design

Recently, some investigators proposed using a multi-stage population-based family study design, which combines features of familial genetic studies (linkage and segregation) and population-based association studies. The case-control family study design is the most important part of the proposed population-based family design. The

investigators suggested that gene-environment interaction in disease etiology can be incorporated in this study design (92).

C. Variance component approach

The variance component approach has been used mainly in the analysis of familial aggregation for quantitative traits (93-94). Recently, investigators have extended the variance component approach to the analysis of dichotomous traits (95) and to a study of the gene-gene (epistasis) interaction in linkage analysis (96). Others have discussed extending this approach to include gene-environment interaction in linkage analysis for both quantitative and qualitative traits (97).

VI Conclusion

In this paper we attempted to provide an overview of both traditional and non-traditional epidemiologic approaches to studying gene-environment interaction. There is little doubt that studies of complex traits in human will assume a central place in the future genetic analysis of common human diseases (40). It is believed that human common diseases are more likely to involve multiple genes with modest effects and gene-environment interaction (37, 40). The modest effects of these genes may indicate that a larger proportion of the disease in the population may be attributed to these genes.

With advances in genetic technology and the work of the Human Genome Project, methods of studying gene-environment interaction will continue to evolve, and the concept of gene-environment interaction will become a central theme in epidemiologic studies that assess causes of human disease in populations.

Acknowledgments

The authors thank Drs James Buehler, Dana W. Flanders and Thomas Sinks for their helpful suggestions. The authors also thank two reviewers for their many constructive comments on an earlier version of the manuscript.

References

1. Khoury MJ. Genetic Epidemiology. In: Rothman KJ, ed. Modern Epidemiology. Boston: Little, Brown and Company, 1997.

2. Koopman JS. Causal models and sources of interaction. Am J Epidemiol 1977;106:439-44.

3. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980;112:467-70.

4. Siemiatycki J, Thomas DC. Biological models and statistical interactions. Int J epidemiol 1981;10:383-87.

5. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. New York: Van Nostrand Reinhold, 1982.

6. Miettinen OS. Causal and preventive interdependence: elementary principles. Scand J Work Environ Health 1982;18:159-68.

7. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983;2:243-51.

8. Rothman KJ. Modern Epidemiology. Boston: Little, Brown, and Company, 1986.

9. Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 1988;14:125-29.

10. Pearce N. Analytical implications of epidemiological concepts of interaction. Int J Epidemiol 1989;18:976-80.

11. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 1991;44:221-32.

12. Greenland S. Basic problems in interaction assessment. Environ Health Perspect 1993;101(suppl 4):59-66.

13. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology 1992;3:452-456.

14. Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. Epidemiology 1996;7:286-290.

15. Blot Wj, Day NE. Synergism and interaction: are they equivalent? Am J Epidemiol 1979;110:99-100.

16. Seminara D, Obrams GI. Genetic epidemiology of cancer: a multidisciplinary approach. Genet Epidemiol 1994;11:235-54.

17. Claus EB, Schildkraut JM, Thompson WD, et al. The genetic attributable risk of breast and ovarian cancer. Cancer

1996;77(11):2318-24.

18. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene: BRCA1. Science 1994;266:66-71.

19. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA1, to chromosome 13q12-13. Science 1994;265:2088-90.

20. Fitzgerald MG, MacDonald DJ, Krainer M, et al. Germ-line BRCA1 mutations in Jewish and non-Jewish women with early-onset breast cancer. New Engl J Med 1996;334:143-49.

21. Dorman JS. Genetic epidemiology of insulin-dependent diabetes mellitus: international comparisons using molecular genetics. Ann Med 1992;24:393-9.

22. Silman AJ. The genetic epidemiology of rheumatoid arthritis. Clin Exp Rheumatol 1992;10:309-12.

23.Van Dujin CM, Clayton DG, Chandra V, et al. Interaction between genetic and environmental risk factors for Alzheimer-s disease: a reanalysis of case-control studies. Genet Epidemiol 1994;11:539-51.

24. Khoury, M.J., Beaty, T.H., Cohen, B.H. Fundamentals of Genetic Epidemiology. New York: Oxford University Press, 1993.

25. Ottman R. Epidemiologic analysis of gene-environment interaction in twins. Genet Epidemiol 1994;11:75-86.

26. Haldane JBS. The interaction of nature and nurture. Annals of Eugenics 1946;13:197-205.

27. Omenn GS, Motulsky AG. Ecogenetics: genetic variation in susceptibility to environmental agents. In: Cohen BH, Lilienfeld AM, Huang PC eds. Genetic Issues in Public Health and Medicine. Springfield IL: CC Thomas, 1978.

28. Calabrese EJ. Ecogenetics: Genetic Variation in Susceptibility to Environmental Agents. New York: Wiley, 1984.

29. Mulvihill JJ. Clinical ecogenetics of cancer in humans. In: Genes and Cancer. New York: Alan R Liss, 1984.

30. Khoury MJ, Adams MJ, Flanders WD. An epidemiologic approach to ecogenetics. Am J Hum Genet 1988;42:89-95.

31. Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990;7:177-85.

32. Ottman R. Gene-environment interaction and public health. Am J Hum Genet 1995;56:821-23.

33. Khoury MJ, Beaty TH. Applications of the case-control method in genetic epidemiology. Epidemiol Rev 1994;16:134-50.

34. Khoury MJ, Flanders WD. Non-traditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996;144:207-13.

35. Cheng TJ, Christiani DC, Xu X, et al. Glutathione S-transferase mu genotype, diet, and smoking as determinants of sister chromatid exchange frequency in lymphocytes. Cancer Epidemiol Biolmarkers Prev 1995; 4(5):535-42.

36. Campbell H. Gene environment interaction. J Epidemiol Community Health 1996;50:397-400.

37. Lander ES. The new genomics: global views of biology. Science 1996;274:536-39.

38. Khoury MJ, Flanders WD. Illustration of the effects of genotype misclassification on the measurement of familial aggregation in epidemiologic studies. Epidemiology 1990;1:51-57.

39. Khoury, M.J., Flanders, W.D. Bias in using family history as a risk factor in case-control studies of disease. Epidemiology 1995;6:511-19.

40. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037-48.

41. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-517.

42. Hwang SJ, Beaty TH, Panny SR, et al. Association study of transforming growth factor alpha TaqI polymorphisms and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. Am J Epidemiol 1995;141:629-36.

43. Yamori Y, Nara Y, Mizushima S, et al. Gene-environment interaction in hypertension, stroke and atherosclerosis in epidemiological survey: a WHO-cardiac study. Clinical & Experimental Pharmacology & Physiology - Supplement. 1992;20:43-52.

44. Susser E, Susser M. Familial aggregation studies: a note on their epidemiologic properties. Am J Epidemiol 1989;129:23-30.

45. Morgenstern H, Thomas D. Principles of study design in environmental epidemiology. Environ Health Perspect Supplements 1993;101(4):23-39.

46. Khoury MJ, Walter S, Beaty TH. The effect of genetic susceptibility on causal inference in epidemiolgic studies. Am J Epidemiol 1987;126:561-67.

47. Morton NE. Linkage and association. Prog Clin Biol Res 1984;147:245-65.

48. Rothman N, Stewart WF, Caporaso NE, Hayes RB. Misclassification of genetic susceptibility biomarkers: implications for case-control studies and cross-population comparisons. Cancer Epidemiology, Biomarkers & Prevention 1993;2:299-303.

49. Vineis P, Schulte PA, Vogt RF. Technical variability in laboratory data. In: PA Schulte and FP Perera (eds.), Molecular Epidemiology: principles and practices, pp. 109-135. San Diego: Academic Press, 1993.

50. Hayes RB, Bi W, Rothman N, et al. N-acetylation phenotype and genotype and risk of bladder cancer in benzidine exposed workers. Carcinogenesis 1993;14:675-78.

51. Rothman N, Hayes RB, Bi W, et al. Correlation between N-acetyltransferase activity and NAT2 genotype in Chinese males. Pharmacogenetics 1993;3(5):250-55.

52. Khoury MJ. Case-parental control method in the search for disease susceptibility genes. Am J Hum Genet 1994;55:414-415.

53. Knowler WC, Williams RC, Pettit DJ, et al. Gm3,5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 1988;43:520-26.

54. Thomas DC. Are dose-response, synergy, and latency confounded? Alexandria, VA: American Statistical Association, 1981.

55. Risch A, Wallace DM, Bathers S, et al. Slow N-acetylation genotype is a susceptibility factor in occupational and smoking related bladder cancer. Hum Mol Genet 1995;4(2):231-36.

56. Ambrosone CB, Freudenheim JL, Graham S, et al. Cigarette smoking, N-acetyltransferase 2 genetic polymorphisms, and breast cancer risk. JAMA 1996;276:1494-512.

57. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984;13:356-65.

58. Wickramaratne PJ. Sample size determination in epidemiologic studies. 1995;4:311-37.

59. Khoury MJ, Beaty TH, Hwang, SJ. Detection of genotype-environment interaction in case-control studies of birth defects: how big a sample size? Teratology 1995;51:336-43.

60. Hwang SJ, Beaty TH, Liang KY, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994;140:1029-37.

61. Andrieu N, Goldstein AM. Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. Int J Epidemiol 1996;25(3):649-57.

62. Struewing JP, Abeliovich D, Peretz T, et al. The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. Nat Genet 1995;11:198-200.

63. Kelsey JL (ed). Breast cancer. Epidemiol Reviews 1993;15:1-263.

64. Weinberg CR, Wacholder S. The design and analysis of case-control studies with biased sampling. Biometrics 1990;46:963-75.

65. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. Am J Epidemiol 1991;134:421-32.

66. Yang Quanhe, Khoury MJ, WD Flanders. Sample size requirments in case-only designs to detect gene-environment interaction. Am J Epidemiol 1997 (in press).

67. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994;13:153-62.

68. Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. Cancer Epidemiol Biomarkers Prev 1994;3:173-75.

69. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus. Am J Hum Genet 1993;52:506-16.

70. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 1995;57:455-64.

71. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59:983-89.

72. Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 1987;51:227-33.

73. Ott J. Statistical properties of the haplotype relative risk. Genet Epidemiol 1989;6:127-30.

74. Knapp M, Seuchter SA, Baur MP. The haplotype relative risk method for analysis of associations in nuclear families. Am J Hum Genet 1993;52:1085-93.

75. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate gene association studies. Am J Hum Genet 1993;53:1114-26.

76. Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and their parents. Am J Hum Genet 1994;55:402-9.

77. Thomson G. Mapping disease genes: family-based association studies. Am J Hum Genet 1995;57:487-98.

78. Flanders WD, Khoury MJ. Analysis of case-parental control studies: method for the study of association between disease and genetic markers. Am J Epidemiol 1996;144(7):696-703.

79. Schaid DJ, Li HZ. Genotype relative-risk and association tests for nuclear families with missing parental data. Genet Epidemiol (in press).

80. Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3-19.

81. Risch N. Linkage strategies for genetically complex traits: II. The power of affected relative pairs. Am J Hum Genet 1990;46:229-41.

82. Fulker DW, Cherry SS, Cardon LR. Multipoint interval mapping of quantitative trait loci using sib pairs. Am J Hum Genet 1995;56:1224-33.

83. Knapp M, Seuchter SA, Baur MP. Linkage analysis in nuclear families. 1. Optimality criteria for affected sib-pair tests. Hum Hered 1994;44:37-43.

84. Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 1995;57:439-54.

85. Khoury MJ, Flanders WD, Lipton RB, et al. The affected sib-pair method in the context of an epidemiologic study design. Genet Epidemiol 1991;8:277-82.

86. Flanders WD, Khoury MJ. Extensions to methods of sib-pair linkage analysis. Genet Epidemiol 1991;8:399-408.

87. Yang Quanhe, Atkinson M, Sun FZ, et al. The method of sib-pair linkage analysis in context of case-control design. Genet Epidemiol (in press).

88. Mayer EJ, Newman B, Austin MA, et al. Genetic and environmental influences on insulin levels and the insulin resistance syndrome: an analysis of women twins. Am J Epidemiol 1996;143:323-32.

89. Ramakrishman V, Goldberg J, Henderson WG, et al. Elementary methods for the analysis of dichotomous

outcomes in unselected samples of twins. Genet Epidemiol 1992;9:273-87.
90. Olson J, Schmidt MM, Christensen K. Evaluation of nature-nurture impact on reproductive health using half-siblings. Epidemiology 1997;8:6-11.

91. Scott WK, Pericak-Vance MA, Haines JL, et al. Genetic analysis of complex diseases. Science 1997;275:1327-30.

92. Hsu L, Davidov O, Holte S, et al. A population based family study of a common oligogenic disease (I): design. Genet Epidemiol (in press).

93. Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 1994;54:535-43.

94. Blangero J. Multivariate oligogenic linkage analysis of quantitative traits in general pedigrees. Am J Hum Genet 1995;57:A11:50.

95. Duggirala R, Williams JT, Williams-Blangero S, et al. A variance component approach to dichotomous trait linkage analysis using a threshold model. Genet Epidemiol (in press).

96. Dupuis J, Brown PO, Siegmund D. Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent. Genetics 1995;140:843-56.

97. Blangero J, Almasy LA. SOLAR: sequential oligogenic linkage analysis routines. Technical notes. San Antonio, TX: Population Genetics Laboratory, Southwest Foundation for Biomedical Research, 1996.

Table 1. A Simple Gene-Environment Interaction Model in the Context of Epidemiologic Studies

Cohort Study				Case-control study		
Exposure (1=present, 0=absent)	**Susceptibility Genotype**	**Disease Risk**	**Relative Risk**	**Cases**	**Controls**	**Odds Ratio**
0	0	I	1	A_{00}	B_{00}	1
0	1	IR_g	R_g	A_{01}	B_{01}	$R_g=A_{01}B_{00}/A_{00}B_{01}$
1	0	IR_e	R_e	A_{10}	B_{10}	$R_e=A_{10}B_{00}/A_{00}B_{10}$
1	1	IR_{ge}	R_{ge}	A_{11}	B_{11}	$R_{ge}=A_{11}B_{00}/A_{00}B_{11}$

I refers to the background disease risk, incidence of disease among members of the cohort
who are not exposed to the environmental factor and who are genotype negative.
R_e = disease risk among persons with the exposure without the genotype divided by disease risk among persons with no exposure and no susceptible genotype.
R_g = disease risk among persons with the genotype without the exposure divided by disease risk among persons with no exposure and no susceptible genotype.
R_{ge} = disease risk among persons with the exposure and genetype divided by disease risk among persons with no exposure and no susceptible genotype.

Table 2. Six Patterns of Gene-Environment Interaction

Patterns	Effects on Disease: Genotype in absence of environment	Risk of: Environment in absence of genotype
1	No effect $R_g = 1$	No effect $R_e = 1$
2	No effect $R_g = 1$	Increase risk $R_e > 1$
3	Increase risk $R_g > 1$	No effect $R_e = 1$
4	Increase risk $R_g > 1$	Increase risk $R_e > 1$
5	Decrease risk $R_g < 1$	No effect $R_e = 1$
6	Decrease risk $R_g < 1$	Increase risk $R_e > 1$

Source: Khoury et al. 1993 (24).

R_e = disease risk among persons with the exposure without the genotype divided by disease risk among persons with no exposure and no susceptible genotype.
R_g = disease risk among persons with the genotype without the exposure divided by disease risk among persons with no exposure and no susceptible genotype.

Table 3. Parameters of Gene-Environment Interaction Analysis in a Case-Control Design

Exposure	Susceptibility Genotype	Cases	Controls	Odds Ratio
-	-	$\frac{(1-g)(1-e)}{3}$	(1-g)(1-e)	1.0
-	+	$\frac{g(1-e)R_g}{3}$	g(1-e)	R_g
+	-	$\frac{e(1-g)R_e}{3}$	e(1-g)	R_e
+	+	$\frac{geR_{ge}}{3}$	ge	R_{ge}

e = prevalence of exposure in the population.
g = prevalence of genotype in the population.
R_e = disease risk among persons with the exposure without the genotype divided by disease risk among persons with no exposure and no susceptible genotype.
R_g = disease risk among persons with the genotype without the exposure divided by disease risk among persons with no exposure and no susceptible genotype.
R_{ge} = disease risk among persons with the exposure and genetype divided by disease risk among persons with no exposure and no susceptible genotype.

$3 = (1-g)(1-e) + g(1-e)R_g + e(1-g)R_e + geR_{ge}$

Table 4. Characteristics of Case-Only, Case-Parental and Affected Sib-pair Studies

Feature	Case-Only	Case-Parental Control	Affected Relative-Pair
Study subjects	Cases	Cases and their parents	Proband, second case in family, and parents
'Controls'	None	Expected genotype distribution based on parental genotypes	Expected distribution of alleles with Mendelian transmission
Assessment	Departure from multiplicative relationship between exposure and genotype	Association between genotype and disease	Linkage between locus and disease
Assumptions	Independence between genotype and exposure	Mendelian transmission	Mendelian transmission
Limitations	Cannot assess effects of exposure on genotype. Linkage disequilibrium.	Requires one or both parents. Cannot assess exposure effects. Linkage disequilibrium.	Need families with 2 or more cases. Cannot assess exposure. Cannot assess specific alleles.

Source: Khoury, 1997 (1)

Table 5. Gene-Environment Interaction Analysis in the Context of a Case-Only Study

Exposure	Susceptibility	Genotype
	-	+
-	a	b
+	c	d

a = ((1-g)(1-e)) / 3
b = ((1-g)eR_e) / 3
c = ((1-e)gR_g) / 3
d = (geR_{ge}) / 3

e = prevalence of exposure in the population.
g = prevalence of genotype in the population.
R_e = disease risk among persons with the exposure without the genotype divided by disease risk among persons with no exposure and no susceptible genotype.
R_g = disease risk among persons with the genotype without the exposure divided by disease risk among persons with no exposure and no susceptible genotype.
R_{ge} = disease risk among persons with the exposure and genetype divided by disease risk among persons with no exposure and no susceptible genotype.

3 = (1-g)(1-e) + g(1-e)R_g + e(1-g)R_e +geR_{ge}

Under assumption of independence between exposure and genotype among controls: case-only odds ratio (OR_{ca})= ad/bc. OR_{ca} is related to case-control ORs by $OR_{ca} = R_{ge}/(R_e*R_g)$.

Table 6. Case-Control Analysis of the Interaction Between Maternal Cigarette Smoking and Transforming Growth Factor Alpha Polymorphism in Determining Children's Risk for Cleft Palate

Smoking	TaqI Polymorphism	Cases	Controls	Odds Ratio	95% C.I.
-	-	36	167	1.0	Referent
-	+	7	34	1.0	0.3-2.4
+	-	13	69	0.9	0.4-1.8
+	+	13	11	5.5	2.1-14.6

Sources: it is derived from Hwang et al. (42).
Odds ratio based on a case-only study is 5.1 (95% CI 1.5-18.5)(36 * 13)/(13 * 7).

Table 7. Gene-Environment Interaction Analysis in the Context of a Case-Parental Control Study: Analysis of Nontransmitted Alleles

Exposure status: Absent		**Case genotype**	
		S	+
Parental non-transmitted alleles	-	T_0	U_0
	+	V_0	W_0
OR among unexposed people		1	U_0/V_0

Exposure status: Present		**Case genotype**	
		S	+
Parental non-transmitted alleles	-	T_1	U_1
	+	V_1	W_1
OR among exposed people		1	U_1/V_1

Source: Khoury and Flanders, 1996 (34).

Table 8. Gene-Environment Interaction Analysis in the Context of an Affected Sib-Pair Study

No. Alleles ibd with proband	**Unexposed case**	**Exposed case**	**Expected**	**Odds Ratio** (unexposed)	**Odds Ratio** (exposed)
0	A_{00}	A_{01}	0.25	1.0	1.0
1	A_{10}	A_{11}	0.50	$A_{10}/2A_{00}$	$A_{11}/2A_{01}$
2	A_{20}	A_{21}	0.25	A_{20}/A_{00} A_{21}/A_{01}	

Source: Khoury, 1997 (1).